Theoretical Computer Science 3 (1976) 321–348.
© North-Holland Publishing Company

ANALYSIS OF A SIMPLE FACTORIZATION ALGORITHM*

Donald E. KNUTH and Luis TRABB PARDO
Computer Science Department, Stanford University, Stanford, CA 94305, U.S.A.

Communicated by M.S. Paterson
Received 30 March 1976
Revised 27 August 1976

Abstract. The probability that the k^{th} largest prime factor of a number n is at most n^x is shown to approach a limit $F_k(x)$ as $n \to \infty$. Several interesting properties of $F_k(x)$ are explored, and numerical tables are given. These results are applied to the analysis of an algorithm commonly used to find all prime factors of a given number. The average number of digits in the k^{th} largest prime factor of a random m-digit number is shown to be asymptotically equivalent to the average length of the k^{th} longest cycle in a permutation on m objects.

0. Introduction

Perhaps the simplest way to discover the prime factorization of an integer n is to try dividing it by $2, 3, 4, 5, \ldots$ and to "cast out" each factor that is discovered; we stop when the trial divisor exceeds the square root of the remaining unfactored part.

The speed of this method obviously depends on the size of the prime factors of n. For example, if n is prime, the number of trial divisions is approximately $n^{1/2}$; but if n is a power of 2, the number is only about $\log n$. In this paper we shall analyze the algorithm when n is a "random" integer, determining the approximate probability that the number of trial divisions is $\leqslant n^x$ when x is a given number between 0 and 1/2. One of the results we shall prove is that the number of trial divisions will be $\leqslant n^{0.35}$, about half of the time.

In order to carry out the analysis, we shall study the distribution of the k^{th} *largest prime factor* of a random integer. This problem is of independent interest in number theory, and for $k > 1$ it does not appear to have been studied before. (Wunderlich and Selfridge [16] gave a heuristic argument that the second-largest prime factor will tend to be roughly $(n^{1-0.61})^{0.61} \approx n^{0.24}$ because the median value of

* This research was supported in part by National Science Foundation grant DCR72-03752 A02, by the Office of Naval Research contract NR 044-402, and by IBM Corporation. Some of the computations and formula manipulation were done with the MACSYMA system supported by the Advanced Research Projects Ageney (ARPA), under Office of Naval Research contract N00014-70-A-0362-0001. Reproduction in whole or in part is permitted for any purpose of the United States Government.

the largest prime factor is $\approx n^{0.61}$; besides their remark, which stimulated the present investigation, the authors are not aware of any published study of the second-largest prime factor. John M. Pollard [private communication] has independently investigated the distribution of second-largest prime factors, and his computed values agree with those presented below.)

Section 1 of this paper presents the factorization algorithm in detail and proves its correctness. Quantitative analysis begins in Section 2, where the two frequency counts involved in the running time are interpreted in terms of the size of the largest two prime factors.

The distribution of k^{th} largest prime factors is investigated heuristically in Section 3, somewhat as a physicist might do the analysis. A rigorous derivation of this distribution, somewhat as a mathematician might do the analysis, is presented in Section 4. Sections 5 and 6 continue the mathematical play by deriving interesting identities and asymptotic formulas satisfied by these distributions. Section 7 comes back to the factorization procedure and applies the ideas to the results of Sections 1 and 2, somewhat as a computer scientist might do the analysis.

Section 8 discusses the particular theoretical model used in these analyses, and expains why the traditional "mean and variance'' approach is inappropriate for algorithms such as this. Numerical tables and empirical confirmation of the theory appear in Section 9. Finally, Section 10 discusses a rather surprising connection between prime factors of random m-digit integers and the cycle lengths of random permutations on m objects.

Although we shall deal with a very simple approach to factoring, the results and methods of this paper apply to many other algorithms as well. The paper is self-contained, and includes several examples suitable for classroom exposition of asymptotic methods.

1. The algorithm

Here is the standard "divide and factor" algorithm which we shall analyze in detail. A proof of its validity follows immediately from the following invariant assertions governing the variables used:

$$n \geqslant 2; \tag{1.1}$$

$$n = p_1 \cdots p_t m; \tag{1.2}$$

$$p_1, \ldots, p_t \quad \text{are prime numbers;} \tag{1.3}$$

$$m \geqslant d; \tag{1.4}$$

$$\text{all prime factors of } m \text{ are } \geqslant d. \tag{1.5}$$

Since our goal is to analyze a simple algorithm rather than to present it in optimized form ready for extensive use, we shall simply consider the following informal Algol-like description:

```
$t:=0$; $m:=n$; $d:=2$;                                        1
while $d^2 \leqslant m$ do                                     $D+1$
begin increase $d$ or decrease $m$:
  if $d$ divides $m$ then                                      $D$
    begin
      $t:=t+1; p_t:=d; m:=m/d$                                 $T-1$
    end
    else $d:=d+1$                                              $D-T+1$
end
$t:=t+1; p_t:=m; m:=1; d:=1$;                                  1
```

The invariant assertions hold after each line of this program. The expressions in the right-hand column specify the number of times the operations in a particular line will be performed, where

$$D \text{ is the number of trial divisions performed,} \tag{1.6}$$

$$T \text{ is the number of prime factors of } n \text{ (counting multiplicity).} \tag{1.7}$$

The usual refinements of this algorithm, which avoid a lot of nonprime trial divisors by making d run through only values of the form $6k \pm 1$, say, when $d > 3$, have the effect of dividing D by a constant; so our analysis of this simple case will apply also with minor variations to the more complicated cases.

2. Preliminary analysis

Let n_k be the k^{th} largest prime factor of n; thus $n_k = p_{T+1-k}$ after the above algorithm terminates, for $1 \leqslant k \leqslant T$. If n has less than k prime factors (counting multiplicities), let $n_k = 1$. We also let $n_0 = \infty$ for convenience in what follows.

The ***while*** loop in the algorithm can terminate in three different ways, depending on how we last encounter it:

Case 1. $n < 4$. Then $D = 0$.

Case 2. $n \geqslant 4$ and the D^{th} trial division succeeds. Then the final trial division was by $d = n_2$, where $d^2 > n_1$. Since d is initially 2 and the statement $d := d+1$ is performed $D - T + 1$ times, we have

$$D = n_2 + T - 3, \qquad n_2^2 > n_1. \tag{2.1}$$

Case 3. $n \geqslant 4$ and the D^{th} trial division fails. Then the final trial division was by d, where $n_2 \leqslant d$ and $d^2 < n_1$ and $(d+1)^2 > n_1$. (Note that if we set $p_0 := 1$ we have $d \geqslant p_{t-1}$ throughout the ***while*** loop.) Thus we have

$$D = \lceil\sqrt{n_1}\rceil + T - 3, \quad n_2^2 < n_1. \tag{2.2}$$

In all three cases we have the formula

$$D = \max(n_2, \lceil\sqrt{n_1}\rceil) + T - 3. \tag{2.3}$$

Clearly D is the dominant factor in the running time, so most of our analysis will be devoted to it. However, it turns out that the analysis of T is also very interesting; for large random n, the number T of prime factors can be regarded as a normally-distributed random variable with mean $\ln\ln n + 1.03$ and standard deviation $\sqrt{\ln\ln n}$ (see Appendix A).

3. The k^{th} largest prime factor

In order to analyze D, we shall first analyze the distributions of n_1 and n_2 (and n_k in general). This analysis will be of interest in itself, since one of the common improvements to the simple algorithm in Section 1 is to test m for primality (by another method) whenever a new factor has been cast out; then the running time for factorization is essentially governed by n_2 alone, rather than $\max(n_2, \sqrt{n_1})$.

Let $P_k(x, N)$ be the number of integers n in the range $1 \leqslant n \leqslant N$ such that

$$n_k \leqslant N^x, \tag{3.1}$$

where x is any number $\geqslant 0$. Thus $P_k(x,N)/N$ is the probability that a random integer between 1 and N will have k^{th} largest prime factor $\leqslant N^x$. We will prove that this probability tends to a limiting distribution

$$\lim_{N\to\infty} \frac{P_k(x,N)}{N} = F_k(x), \tag{3.2}$$

where $F_k(x)$ has interesting properties discussed below.

Before we establish (3.2) rigorously, it will be helpful to give a heuristic derivation analogous to that given by Dickman [4], who was the first to study this question in the case $k = 1$. Let us consider $P_k(t + \mathrm{d}t, N) - P_k(t, N)$, the number of $n \leqslant N$ such that n_k lies between N^t and $N^{t+\mathrm{d}t}$, when $\mathrm{d}t$ is very small. To count the number of such n, we take all primes p lying between N^t and $N^{t+\mathrm{d}t}$, and multiply by all numbers $m \leqslant N^{1-t}$ such that $m_k \leqslant p$ and $m_{k-1} \geqslant p$. Now if $n = mp$ we have $n \leqslant N^{t+\mathrm{d}t}$ and $n_k = p$; conversely every $n \leqslant N$ with n_k between N^t and $N^{t+\mathrm{d}t}$ will have the form $n = mp$ where p and m have the stated form. Note that the number of $m \leqslant N^{1-t}$ such that $m_k \leqslant p$ is approximately $P_k(t/(1-t), N^{1-t})$, and the unwanted subset consisting of those m with $m_{k-1} < p$ has approximately $P_{k-1}(t/(1-t), N^{1-t})$ members. Hence the number of m with $mp \leqslant N$ and $m_k \leqslant p$ and $m_{k-1} \geqslant p$ is $P_k(t/(1-t), N^{1-t}) - P_{k-1}(t/(1-t), N^{1-t})$, ignoring second-order terms, and we have

$$P_k(t+\mathrm{d}t, N) - P_k(t, N)$$
$$\approx (\pi(N^{t+\mathrm{d}t}) - \pi(N^t))(P_k(t/(1-t), N^{1-t}) - P_{k-1}(t/(1-t), N^{1-t}). \tag{3.3}$$

Here the π function is defined as usual,

$$\pi(x) = \text{the number of primes not exceeding } x. \tag{3.4}$$

According to the prime number theorem we have $\pi(x) \approx x/\ln x$, hence

$$\pi(N^{t+\mathrm{d}t}) - \pi(N^t) \approx N^t\,\mathrm{d}t/t. \tag{3.5}$$

Plugging this into the above formula and dividing by N yields

$$\frac{P_k(t+\mathrm{d}t, N) - P_k(t, N)}{N} \approx \frac{\mathrm{d}t}{t}\left(\frac{P_k(t/(1-t), N^{1-t})}{N^{1-t}} - \frac{P_{k-1}(t/(1-t), N^{1-t})}{N^{1-t}}\right); \tag{3.6}$$

when $N \to \infty$ we have the differential equation

$$F_k'(t)\mathrm{d}t = \frac{\mathrm{d}t}{t}\left(F_k\left(\frac{t}{1-t}\right) - F_{k-1}\left(\frac{t}{1-t}\right)\right). \tag{3.7}$$

Since $F_k(0) = 0$, we may integrate (3.7) to deduce the formula

$$F_k(x) = \int_0^x \left(F_k\left(\frac{t}{1-t}\right) - F_{k-1}\left(\frac{t}{1-t}\right)\right)\frac{\mathrm{d}t}{t}. \tag{3.8}$$

According to our convention $n_0 = \infty$, we define

$$F_0(x) = 0 \quad \text{for all } x. \tag{3.9}$$

We also must have

$$F_k(x) = 1 \quad \text{for } x \geq 1,\ k \geq 1. \tag{3.10}$$

Now it is easy to see that (3.8), (3.9), (3.10) define $F_k(x)$ uniquely for $0 \leq x \leq 1$, since we have

$$F_k(x) = 1 - \int_x^1 \frac{\mathrm{d}t}{t}\left(F_k\left(\frac{t}{1-t}\right) - F_{k-1}\left(\frac{t}{1-t}\right)\right), \qquad 0 \leq x \leq 1 \tag{3.11}$$

and this relation defines $F_k(x)$ in terms of its values at points $> x$.

4. Proof without handwaving

Our discussion in the previous section has been only quasi-rigorous, but it shows that *if* the limiting relationship (3.2) holds then $F_k(x)$ had better be the function defined by (3.8), (3.9), and (3.10). Now that we have a formula for F_k, let us try to prove the limiting formula (3.2).

It is more convenient to work with the functions ρ_k defined by

$$\rho_k(\alpha) = F_k(1/\alpha); \tag{4.1}$$

the above equations transform into the somewhat simpler recurrrence formulas

$$\rho_k(\alpha) = 1 - \int_1^\alpha (\rho_k(t-1) - \rho_{k-1}(t-1))\,\frac{dt}{t}, \quad \text{for } \alpha > 1,\ k \geqslant 1; \tag{4.2}$$

$$\rho_k(\alpha) = 1 \quad \text{for } 0 < \alpha \leqslant 1,\ k \geqslant 1; \tag{4.3}$$

$$\rho_k(\alpha) = 0 \quad \text{for } \alpha \leqslant 0 \text{ or } k = 0. \tag{4.4}$$

Furthermore we let $S_k(x, y)$ be the set of positive integers $n \leqslant x$ such that $n_k \leqslant y$, and let $\Psi_k(x, y) = \|S_k(x, y)\|$ be its cardinality, so that

$$P_k(x, N) = \Psi_k(N, N^x). \tag{4.5}$$

We will show that

$$\Psi_k(N^\alpha, N) = \rho_k(\alpha)N^\alpha + O(N^\alpha/\log N^\alpha), \tag{4.6}$$

and it follows that a stronger form of (3.2) is true:

$$\frac{P_k(x, N)}{N} = F_k(X) + O\left(\frac{1}{\log N}\right). \tag{4.7}$$

Indeed, we will prove a result even stronger than (4.6), namely

$$\Psi_k(x^\alpha, x) = \rho_k(\alpha)x^\alpha + \sigma_k(\alpha)x^\alpha/\ln x^\alpha + O(x^\alpha/(\log x)^2) \tag{4.8}$$

as $x \to \infty$, for all fixed $\alpha > 1$, where $\sigma_k(\alpha)$ will be defined appropriately below. In principle, the approach we shall use could be extended to obtain an asymptotic formula for $\Psi_k(x^\alpha, x)$ which is good to $O(x^\alpha/(\log x)^r)$ for any fixed r; the method is based on ideas of de Bruijn [1], who went on to find extremely precise asymptotic expansions of $\Psi_1(N^\alpha, N)$ in an elegant way using Stieltjes integration by parts. (Note: When $k = 1$, the limiting formula (3.2) was first established by Ramaswami [12]; Norton [10] has given a comprehensive survey of the literature relating to this important special case.)

We shall use a strong form of the prime number theorem due to de la Vallée Poussin [3]:

$$\pi(x) = L(x) + O(x\,e^{-C\sqrt{\log x}}), \tag{4.9}$$

where C is a positive constant and

$$L(x) = \int_2^x \frac{dt}{\ln t}. \tag{4.10}$$

Now to the proof, which will be "elementary" except for our use of (4.9). Letting p range over primes and n over positive integers, we have

$$\begin{aligned}\lfloor x^\alpha \rfloor - \Psi_k(x^\alpha, x) &= \sum_{x<p\leqslant x^\alpha} \|\{n \leqslant x^\alpha \mid n_k = p\}\| \\ &= \sum_{x<p\leqslant x^\alpha} \|\{m \leqslant x^\alpha/p \mid m_k \leqslant p \text{ and } m_{k-1} \geqslant p\}\| \\ &= \sum_{x<p\leqslant x^\alpha} (\Psi_k(x^\alpha/p, p) - \Psi_{k-1}(x^\alpha/p, p-\varepsilon))\end{aligned}$$

where ε is a small positive number and $\Psi_0(x, y) = 0$. The key idea in our derivation will be to replace $\sum_{x<p\leqslant x^\alpha} \Psi_k(x^\alpha/p, p)$ by $\int_x^{x^\alpha} \Psi_k(x^\alpha/y, y)\mathrm{d}y/(\ln y)$, using the "density" function for primes suggested by (4.10). To justify this, we have

$$\begin{aligned}&\Big(\sum_{x<p\leqslant x^\alpha} \Psi_k\left(\frac{x^\alpha}{p}, p\right)\Big) - \int_x^{x^\alpha} \Psi_k\left(\frac{x^\alpha}{y}, y\right)\frac{\mathrm{d}y}{\ln y} \\ &= \Big(\sum_{x<p\leqslant x^\alpha} \sum_{n\in S_k(x^\alpha/p, p)} 1\Big) - \int_x^{x^\alpha}\Big(\sum_{n\in S_k(x^\alpha/y, y)} 1\Big)\frac{\mathrm{d}y}{\ln y} \\ &= \sum_{\substack{1\leqslant n\leqslant x^{\alpha-1}\\ n_k\leqslant x^\alpha/n}} \left(\Big(\sum_{\substack{n_k\leqslant p\leqslant x^\alpha/n\\ x<p}} 1\Big) - \int_{\max(n_k, x)}^{x^\alpha/n} \frac{\mathrm{d}y}{\ln y}\right) \\ &= \sum_{\substack{1\leqslant n\leqslant x^{\alpha-1}\\ n_k\leqslant x^\alpha/n}} (\pi(x^\alpha/n) - \pi(\max(n_k, x)) + \mathrm{O}(1) - L(x^\alpha/n) + L(\max(n_k, x))) \\ &= \sum_{\substack{1\leqslant n\leqslant x^{\alpha-1}\\ n_k\leqslant x^\alpha/n}} \mathrm{O}\left(\frac{x^\alpha}{n}\,\mathrm{e}^{-C\sqrt{\log x}}\right) \\ &= \mathrm{O}(x^\alpha(\log x^\alpha)\mathrm{e}^{-\sqrt{\log x}}).\end{aligned} \tag{4.11}$$

A similar estimate applies to $\sum_{x<p\leqslant x^\alpha} \Psi_{k-1}(x^\alpha/p, p-\varepsilon)$, so we have

$$\Psi_k(x^\alpha, x) = x^\alpha - \int_x^{x^\alpha}\left(\Psi_k\left(\frac{x^\alpha}{y}, y\right) - \Psi_{k-1}\left(\frac{x^\alpha}{y}, y\right)\right)\frac{\mathrm{d}y}{\ln y} + \mathrm{O}\left(\frac{\alpha x^\alpha}{(\log x)^r}\right) \tag{4.12}$$

as $x \to \infty$, for all fixed $r \geqslant 0$. This is the formula we shall use for $\alpha \geqslant 1$; for $0 \leqslant \alpha < 1$ we have $\Psi_k(x^\alpha, x) = \lfloor x^\alpha \rfloor$. (The brackets $\lfloor\ \rfloor$ in the latter formula turn out to be important, since the integral (4.12) is sensitive to O(1) terms in the vicinity of $y = x^\alpha$.)

Our proof of (4.8) is by induction on k, and for fixed k by induction on $\lceil\alpha\rceil$. Actually the first case $k = 1$, $\lceil\alpha\rceil = 2$ seems to be the hardest; when $1 < \alpha \leqslant 2$ we have

$$\begin{aligned}\Psi_1(x^\alpha, x) &= x^\alpha - \int_x^{x^\alpha} \Psi_1\left(\frac{x^\alpha}{y}, y\right)\frac{\mathrm{d}y}{\ln y} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right) \\ &= x^\alpha - \int_x^{x^\alpha}\left(\frac{x^\alpha}{y} - \left\{\frac{x^\alpha}{y}\right\}\right)\frac{\mathrm{d}y}{\ln y} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right) \\ &= x^\alpha - x^\alpha \ln\alpha + x^\alpha\int_1^{x^{\alpha-1}} \{u\}\frac{\mathrm{d}u}{u^2 \ln x^\alpha/u} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)\end{aligned}$$

$$= x^\alpha\rho_1(\alpha) + \frac{x^\alpha}{\ln x^\alpha}\int_1^{x^{\alpha-1}}\left(\frac{\{u\}\mathrm{d}u}{u^2} + \frac{\{u\}\ln u\,\mathrm{d}u}{u^2\ln(x^\alpha/u)}\right) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)$$

$$= x^\alpha\rho_1(\alpha) + \frac{x^\alpha}{\ln x^\alpha}\int_1^{\infty}\frac{\{u\}\mathrm{d}u}{u^2} + \mathrm{O}\left(\frac{x}{\log x}\right) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right), \tag{4.13}$$

where $\{x\}$ denotes $x - \lfloor x \rfloor$. The remaining integral is

$$\int_1^\infty \frac{\{u\}\mathrm{d}u}{u^2} = \sum_{n\geqslant 1}\int_n^{n+1}\frac{(u-n)\mathrm{d}u}{u^2} = \sum_{n\geqslant 1}\left(\left(\ln\frac{n+1}{n}\right) - \frac{1}{n+1}\right)$$

$$= \lim_{n\to\infty}((\ln n) - (H_n - 1)) = 1 - \gamma, \tag{4.14}$$

where γ is Euler's constant.

Now suppose we have proved that

$$\Psi_1(x^\alpha, x) = x^\alpha\rho_1(\alpha) + (1-\gamma)\frac{x^\alpha}{\ln x^\alpha}\rho_1(\alpha-1) + \mathrm{O}\left(\frac{x}{\log x}\right) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right) \tag{4.15}$$

for $1 < \alpha \leqslant m$, where the bounding constants implied by the O's depend on m but not on x or α. The discussion in the previous paragraph extablishes (4.15) for $m = 2$. We can extend it to the next case by analyzing its value for $m < \alpha \leqslant m+1$:

$$\Psi_1(x^\alpha, x) = x^\alpha - \int_x^{x^\alpha}\Psi_1\left(\frac{x^\alpha}{y}, y\right)\frac{\mathrm{d}y}{\ln y} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)$$

$$= x^\alpha - \int_1^\alpha \Psi_1(x^{\alpha(t-1)/t}, x^{\alpha/t})x^{\alpha/t}\frac{\mathrm{d}t}{t} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)$$

$$= x^\alpha - \int_1^2 \lfloor x^{\alpha(t-1)/t}\rfloor x^{\alpha/t}\frac{\mathrm{d}t}{t}$$

$$- x^\alpha\int_2^\alpha\left(\rho_1(t-1) + \frac{(1-\gamma)\rho_1(t-2)}{\ln x^{\alpha(t-1)/t}} + \mathrm{O}\left(\frac{1}{(\alpha/t)\log x}\right)^2\right)\frac{\mathrm{d}t}{t}$$

$$- \int_2^\alpha x^{\alpha/t}\,\mathrm{O}\left(\frac{x^{\alpha/t}}{(\alpha/t)\log x}\right)\frac{\mathrm{d}t}{t} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right) \tag{4.16}$$

by substituting $x^{\alpha/t}$ for y and inserting (4.15). Continuing, we get

$$\Psi_1(x^\alpha, x) = x^\alpha - x^\alpha\int_1^\alpha \rho_1(t-1)\frac{\mathrm{d}t}{t} + \int_1^2\{x^{\alpha(t-1)/t}\}x^{\alpha/t}\frac{\mathrm{d}t}{t}$$

$$-\frac{(1-\gamma)x^\alpha}{\ln x^\alpha}\int_2^\alpha \rho_1(t-2)\frac{\mathrm{d}t}{t-1} + \mathrm{O}\left(\frac{1}{\log x}\int_2^\alpha x^{2\alpha/t}\mathrm{d}t\right) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)$$

$$= x^\alpha\rho_1(\alpha) + x^\alpha\int_1^{x^{\alpha/2}}\frac{\{u\}\mathrm{d}u}{u^2\ln(x^\alpha/u)} - \frac{(1-\gamma)x^\alpha}{\ln x^\alpha}\int_1^{\alpha-1}\rho_1(t-1)\frac{\mathrm{d}t}{t}$$

$$+ \mathrm{O}\left(\frac{1}{\log x}\int_{2/\alpha}^1\frac{x^{\alpha u}\mathrm{d}u}{u^2}\right) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)$$

$$= x^\alpha\rho_1(\alpha) + \frac{x^\alpha(1-\gamma)}{\ln x^\alpha}\rho_1(\alpha-1) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right), \tag{4.17}$$

since

$$\int_1^{x^{\alpha/2}} \frac{\{u\}\mathrm{d}u}{u^2 \ln(x^\alpha/u)} = \frac{1}{\ln x^\alpha}\left(1-\gamma+\mathrm{O}\left(\frac{1}{\log x}\right)\right) \tag{4.18}$$

as in (4.13) and (4.14), and

$$\int_{2/\alpha}^{1} \frac{x^{\alpha u}\mathrm{d}u}{u^2} = \mathrm{O}\left(\int_{2/\alpha}^{1} x^{\alpha u}\mathrm{d}u\right) = \mathrm{O}\left(\frac{x^\alpha}{\log x}\right) \tag{4.19}$$

with bounding constants depending only on m. This establishes (4.15) for all m, by induction.

We have proved (4.8) for $k = 1$, with

$$\sigma_1(\alpha) = (1-\gamma)\rho_1(\alpha-1).$$

For larger k, a similar but simpler derivation applies: Assuming that

$$\Psi_k(x^\alpha, x) = x^\alpha \rho_k(\alpha) + \frac{x^\alpha}{\ln x^\alpha}\,\sigma_k(\alpha) + \mathrm{O}\left(\frac{x}{\log x}\right) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right) \tag{4.20}$$

for $1 < \alpha \leqslant m$ (cf. (4.15)), we extend this to $m < \alpha \leqslant m+1$ by

$$\begin{aligned}
\Psi_k(x^\alpha, x) &= x^\alpha - \int_x^{x^\alpha}\left(\Psi_k\left(\frac{x^\alpha}{y}, y\right) - \Psi_{k-1}\left(\frac{x^\alpha}{y}, y\right)\right)\frac{\mathrm{d}y}{\ln y} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)\\
&= x^\alpha - \int_1^\alpha (\Psi_k(x^{\alpha(t-1)/t}, x^{\alpha/t}) - \Psi_{k-1}(x^{\alpha(t-1)/t}, x^{\alpha/t}))x^{\alpha/t}\frac{\mathrm{d}t}{t} + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right)\\
&= x^\alpha\Bigg(1 - \int_2^\alpha (\rho_k(t-1) - \rho_{k-1}(t-1))\frac{\mathrm{d}t}{t}\\
&\qquad - \frac{1}{\ln x^\alpha}\int_2^\alpha (\sigma_k(t-1) - \sigma_{k-1}(t-1))\frac{\mathrm{d}t}{t-1}\Bigg) + \mathrm{O}\left(\frac{x^\alpha}{(\log x)^2}\right);
\end{aligned} \tag{4.21}$$

the desired relation follows for $k \geqslant 2$ provided that we define

$$\sigma_k(\alpha) = -\int_2^\alpha (\sigma_k(t-1) - \sigma_{k-1}(t-1))\frac{\mathrm{d}t}{t-1} \quad \text{for } \alpha \geqslant 2; \tag{4.22}$$

$$\sigma_k(\alpha) = 0 \quad \text{for } \alpha < 2. \tag{4.23}$$

It follows that

$$\sigma_k(\alpha) = (1-\gamma)(\rho_k(\alpha-1) - \rho_{k-1}(\alpha-1)) \tag{4.24}$$

for all $k \geqslant 1$.

5. Identities satisfied by ρ_k

The functions $\rho_k(\alpha)$ defined by (4.2), (4.3), (4.4) possess many rather surprising properties, and we shall examine some of them in this section.

Our first goal is to express the ρ_k in terms of the polylogarithm functions L_k, defined by

$$L_0(\alpha)=0 \quad \text{for } \alpha \leqslant 0,\ L_0(\alpha)=1 \quad \text{for } \alpha > 0; \tag{5.1}$$

$$L_k(\alpha)=\int_1^\alpha L_{k-1}(t-1)\,\frac{\mathrm{d}t}{t}. \tag{5.2}$$

Thus $L_1(\alpha)=\ln\alpha$ for $\alpha\geqslant 1$, and $L_2(\alpha)=\int_2^\alpha \ln(t-1)\mathrm{d}t/t$ for $\alpha\geqslant 2$, etc.; it is not difficult to verify that $L_k(\alpha)$ is $1/k!$ times the integral of $(\mathrm{d}x_1\cdots\mathrm{d}x_k)/(x_1\cdots x_k)$ over all points $x_1,\ldots,x_k$ where $1\leqslant x_1,\ldots,x_k\leqslant\alpha$ and $|x_i-x_j|\geqslant 1$ for all $i\neq j$. In particular, $L_k(\alpha)=0$ for $\alpha\leqslant k$.

By iterating the recurrence for ρ_k we find

$$1-\rho_1(\alpha)=L_1(\alpha)-L_2(\alpha)+L_3(\alpha)-L_4(\alpha)+L_5(\alpha)-\cdots, \tag{5.3}$$

$$1-\rho_2(\alpha)= \qquad L_2(\alpha)-2L_3(\alpha)+3L_4(\alpha)-4L_5(\alpha)+\cdots, \tag{5.4}$$

for $\alpha>0$, and in general

$$1-\rho_k(\alpha)=\sum_{n\geqslant 0}\binom{-k}{n}L_{n+k}(\alpha). \tag{5.5}$$

These infinite sums are actually finite for any particular value of α.

Now let us examine several auxiliary functions:

$$S_k(\alpha,\beta)=\int_0^\alpha \frac{\rho_k(t-1)\mathrm{d}t}{\beta-t} \quad \text{for } \beta>\alpha \text{ or } \beta<0; \tag{5.6}$$

$$S_k(\alpha)=S_k(\alpha,\alpha+1); \tag{5.7}$$

$$I_k(\alpha)=\int_0^\alpha \frac{\rho_k(t-1)}{t}\ln(t+1)\mathrm{d}t; \tag{5.8}$$

$$\sigma_k(\alpha)=\int_0^\alpha \frac{\rho_k(t-1)}{t}\,\mathrm{d}t; \tag{5.9}$$

$$e_k(x)=\int_0^\infty \rho_k(t)\mathrm{e}^{-tx}\,\mathrm{d}t, \quad x>0. \tag{5.10}$$

(Eq. 5.9 defines a different function $\sigma_k(\alpha)$ from that in Section 4.) It follows immediately from the definition $\rho_k(\alpha)=1-\sigma_k(\alpha)+\sigma_{k-1}(\alpha)$ that

$$\sigma_k(\alpha)=k-\rho_1(\alpha)-\cdots-\rho_k(\alpha). \tag{5.11}$$

Integration by parts enables us to evaluate $I_k(\alpha)$ as follows:

$$I_k(\alpha)-I_{k-1}(\alpha)=-\rho_k(t)\ln(t+1)\Big|_0^\alpha+\int_0^\alpha\frac{\rho_k(t)\mathrm{d}t}{t+1}$$

$$=-\rho_k(\alpha)\ln(\alpha+1)+\sigma_k(\alpha+1). \tag{5.12}$$

Thus in particular we have

$$I_1(\alpha)=-\rho_1(\alpha)\ln(\alpha+1)+1-\rho_1(\alpha+1), \tag{5.13}$$

$$I_2(\alpha)=-\rho_1(\alpha)\ln(\alpha+1)-\rho_2(\alpha)\ln(\alpha+1)+3-2\rho_1(\alpha+1)-\rho_2(\alpha+1), \tag{5.14}$$

etc. A somewhat surprising consequence of this relation is that $I_k(\infty)=k(k+1)/2$, while $\sigma_k(\infty)=k$; in particular, $I_1(\infty)=\sigma_1(\infty)$.

Integration by parts applied to $S_k(\alpha,\beta)$ yields

$$S_k(\alpha,\beta)-S_{k-1}(\alpha,\beta)=-\frac{t\rho_k(t)}{\beta-t}\Big|_0^\alpha+\beta\int_0^\alpha\frac{\rho_k(t)\mathrm{d}t}{(\beta-t)^2}$$

$$=-\frac{\alpha\rho_k(\alpha)}{\beta-\alpha}+\beta\int_1^{\alpha+1}\frac{\rho_k(t-1)\mathrm{d}t}{(\beta+1-t)^2}. \tag{5.15}$$

Differentiating the integral which defines $S_k(\alpha)=S_k(\alpha,\alpha+1)$ with respect to α leads to a formula which can be combined with this one:

$$S_k'(\alpha)=\rho_k(\alpha-1)-\int_1^\alpha\frac{\rho_k(t-1)\mathrm{d}t}{(\alpha+1-t)^2}$$

$$=\rho_k(\alpha-1)-\frac{1}{\alpha}((\alpha-1)\rho_k(\alpha-1)+S_k(\alpha-1)-S_{k-1}(\alpha-1))$$

$$=\frac{1}{\alpha}(\rho_k(\alpha-1)+S_{k-1}(\alpha-1)-S_k(\alpha-1)). \tag{5.16}$$

Now we are ready to prove an important relation which expresses ρ_{k+1} in terms of ρ_k and ρ_{k-1}:

Lemma 5.1.

$$\rho_{k+1}(\alpha)=\rho_k(\alpha)+\frac{1}{k}(S_k(\alpha)-S_{k-1}(\alpha)),\quad \textit{for } k\geqslant 1. \tag{5.17}$$

Proof. Since $\rho_{k+1}(\alpha)=\rho_k(\alpha)=1$ and $S_k(\alpha)=S_{k-1}(\alpha)=0$ for $0<\alpha\leqslant 1$, the result holds for $\lceil\alpha\rceil=1$; we will show that the derivatives agree, by induction on $\lceil\alpha\rceil$. Since

$$(\alpha+1)\rho_{k+1}'(\alpha+1)=\rho_k(\alpha)-\rho_{k+1}(\alpha)=(S_{k-1}(\alpha)-S_k(\alpha))/k,$$

$$(\alpha+1)\rho_k'(\alpha+1)=\rho_{k-1}(\alpha)-\rho_k(\alpha),$$

$$(\alpha+1)S_k'(\alpha+1)=\rho_k(\alpha)+S_{k-1}(\alpha)-S_k(\alpha),$$

$$(\alpha+1)S_{k+1}'(\alpha+1)=\rho_{k-1}(\alpha)+S_{k-2}(\alpha)-S_{k-1}(\alpha),$$

the desired result is equivalent to

$$\frac{k-1}{k}\,\rho_k(\alpha)=\frac{k-1}{k}\,\rho_{k-1}(\alpha)+\frac{1}{k}\,(S_{k-1}(\alpha)-S_{k-2}(\alpha)).$$

For $k=1$ this is obvious, otherwise it holds by induction. □

By iterating the recurrence in the lemma, it follows that

$$\rho_{k+1}(\alpha)=\rho_1(\alpha)+\frac{1}{2.1}\,S_1(\alpha)+\cdots+\frac{1}{k(k-1)}\,S_{k-1}(\alpha)+\frac{1}{k}\,S_k(\alpha). \qquad (5.18)$$

Finally let us consider the functions $e_k(x)$ defined in (5.10). Somewhat surprisingly, these can actually be expressed in closed form:

Theorem 5.2. *$e_k(x)=\mathrm{e}^{-E(x)}x^{-1}(1+E(x)/1!+\cdots+E(x)^{k-1}/(k-1)!)$, where $E(x)=E_1(x)$ is the exponential integral function*

$$E(x)=\int_x^\infty \mathrm{e}^{-t}\mathrm{d}t/t=\int_1^\infty \mathrm{e}^{-xt}\mathrm{d}t/t. \qquad (5.19)$$

Proof. Once again we integrate by parts:

$$\begin{aligned}
e_k(x)-e_{k-1}(x) &= \int_1^\infty \frac{\rho_k(t-1)-\rho_{k-1}(t-1)}{t}\,t\,\mathrm{e}^{-(t-1)x}\,\mathrm{d}t\\
&= -\,\mathrm{e}^x\int_0^\infty t\,\mathrm{e}^{-tx}\,\mathrm{d}\rho_k(t)\\
&= \mathrm{e}^x\int_0^\infty \rho_k(t)(\mathrm{e}^{-tx}-tx\,\mathrm{e}^{-tx})\mathrm{d}t\\
&= \mathrm{e}^x(e_k(x)+xe'_k(x)).
\end{aligned}$$

If we let $f_k(x)=x\mathrm{e}^{E(x)}e_k(x)$, we have therefore

$$\begin{aligned}
f'_k(x) &= \mathrm{e}^{E(x)}(e_k(x)+xe'_k(x)-\mathrm{e}^{-x}e_k(x))\\
&= -\frac{\mathrm{e}^{-x}}{x}\,f_{k-1}(x)=E'(x)f_{k-1}(x)
\end{aligned}$$

and it follows by induction on k that

$$f_k(x)=C+\frac{E(x)}{1!}+\cdots+\frac{E(x)^{k-1}}{(k-1)!}\,.$$

In order to evaluate C, we integrate by parts in the opposite direction:

$$\begin{aligned}
xe_k(x) &= -\int_0^\infty \rho_k(t)\mathrm{d}(\mathrm{e}^{-tx})=-\rho(t)\mathrm{e}^{-tx}\Big|_0^\infty+\int_0^\infty \mathrm{e}^{-tx}\,\mathrm{d}\rho_k(t)\\
&= 1-\int_1^\infty \mathrm{e}^{-tx}(\rho_k(t-1)-\rho_{k-1}(t-1))\,\frac{\mathrm{d}t}{t}\\
&= 1-\int_x^\infty \mathrm{e}^{-u}\left(\rho_k\left(\frac{u}{x}-1\right)-\rho_{k-1}\left(\frac{u}{x}-1\right)\right)\frac{\mathrm{d}u}{u}\,.
\end{aligned}$$

Hence $C = \lim_{x\to\infty} xe_k(x) = \lim_{x\to\infty} f_k(x) = 1$. □

The above theorem now allows us to write down an "explicit" equation for $\rho_k(x)$, instead of a recurrence relation, namely

$$\rho_k(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ixt} e_k(it)\mathrm{d}t, \quad \text{for } x > 0. \tag{5.20}$$

For the identity in the theorem can be extended to all complex x not on the negative real axis, by analytic continuation; then (5.20) follows by fourier inversion. (Note that $E(x) = \ln 1/x + O(1)$ as $x \to 0$, hence $e_k(x) = O(\ln(1/x))^{k-1}$ and the integral is convergent at $t = 0$.) The case $k = 1$ of (5.20) was obtained by de Bruijn in [2].

6. Asymptotic formulas

In this section we shall study the asymptotic behavior of $\rho_k(\alpha)$ for large α. Our starting point is a simple proof that $\rho_1(\alpha)$ is exponentially small: Let us write $\rho(\alpha)$ for $\rho_1(\alpha)$. Then since

$$\begin{aligned} 1 + \int_2^\alpha \rho(t-1)\mathrm{d}t &= \int_1^\alpha \rho(t-1)\mathrm{d}t \\ &= -t\rho(t)\Big|_1^\alpha + \int_1^\alpha \rho(t)\mathrm{d}t \\ &= 1 - \alpha\rho(\alpha) + \int_2^{\alpha+1} \rho(t-1)\mathrm{d}t \end{aligned} \tag{6.1}$$

we have

$$\int_\alpha^{\alpha+1} \rho(t-1)\mathrm{d}t = \alpha\rho(\alpha). \tag{6.2}$$

It follows immediately that $\alpha\rho(\alpha) < \rho(\alpha - 1)$ for all $\alpha > 1$, hence by induction

$$\rho(n) \leqslant 1/n! \tag{6.3}$$

for all integers $n \geqslant 1$. Considerably more precise formulas have been obtained by de Bruijn [1] and others, and numerical results have been tabulated by Mitchell [9] and by van de Lune and Wattel [14]; but (6.3) suffices for our purposes in this section.

The rapid decrease of $\rho_1(\alpha)$ simplifies the numerical evaluation of integrals and it also leads to a simple treatment of the asymptotic behavior of $\rho_2(\alpha)$:

Theorem 6.1. *For all fixed $r \geqslant 1$ we have*

$$\rho_2(\alpha) = A\left(\frac{c_0}{\alpha} + \frac{c_1}{\alpha^2} + \cdots + \frac{c_{r-1}}{\alpha^r}\right) + O(\alpha^{-r-1}) \tag{6.4}$$

as $\alpha \to \infty$, where

$$\mathrm{A} = \mathrm{e}^{\gamma} \approx 1.78107\ 24179\ 90197\ 98524, \tag{6.5}$$

and the coefficients c_k are defined by

$$\sum_{k \geqslant 0} z^k c_k / k! = \exp\left(\int_0^z (\mathrm{e}^t - 1)\mathrm{d}t/t\right) = \exp\left(\sum_{k \geqslant 1} z^k / k \cdot k!\right). \tag{6.6}$$

Thus

$$\langle c_0, c_1, c_2, \ldots \rangle = \left\langle 1, 1, \frac{3}{2}, \frac{17}{6}, \frac{19}{3}, \frac{81}{5}, \frac{8351}{180}, \frac{184553}{1260}, \frac{52907}{105}, \frac{1768847}{945}, \ldots \right\rangle.$$

Before proving the theorem, we note that (6.6) implies the recurrence formula

$$c_n = \frac{1}{n} \sum_{1 \leqslant k \leqslant n} \binom{n}{k} c_{n-k}, \quad n \geqslant 1. \tag{6.7}$$

Therefore $c_n > ((n-1)/2)c_{n-2}$ for $n \geqslant 2$, and $c_{2n+1} > n!$; the infinite series $\sum c_k/\alpha^k$ diverges for all α. In other words, (6.4) is strictly an asymptotic formula.

Proof. From Lemma 5.1, we have

$$\begin{aligned}
\rho_2(\alpha) &= \rho_1(\alpha) + S_1(\alpha) = \rho(\alpha) + \int_1^\alpha \frac{\rho(t-1)\mathrm{d}t}{\alpha + 1 - t} \\
&= \rho(\alpha) + \int_1^\alpha \rho(t-1)\mathrm{d}t \left(\frac{1}{\alpha} + \frac{t-1}{\alpha^2} + \cdots + \frac{(t-1)^r}{\alpha^{r+1}} + \frac{(t-1)^{r+1}}{\alpha^{r+1}(\alpha + 1 - t)}\right) \\
&= \sum_{0 \leqslant k \leqslant r} \int_0^{\alpha - 1} \rho(t) t^k \mathrm{d}t / \alpha^{k+1} + \mathrm{O}(\alpha^{-r-1})
\end{aligned} \tag{6.8}$$

since $\int_1^\alpha \rho(t-1)(t-1)^{r+1}\mathrm{d}t/(\alpha + 1 - t) < \int_1^\infty \rho(t-1)(t-1)^{r+1}\mathrm{d}t < \infty$. Furthermore we have

$$\int_{\alpha-1}^\infty \rho(t) t^k \mathrm{d}t = \mathrm{O}\left(\int_{\alpha-1}^\infty \mathrm{e}^{-t} t^k \mathrm{d}t\right) = \mathrm{O}(\mathrm{e}^{-\alpha/2}) \tag{6.9}$$

as $\alpha \to \infty$, by making very crude estimates not even as powerful as (6.3), so we can integrate to ∞ in (6.8):

$$\rho_2(\alpha) = \frac{a_0}{\alpha} + \frac{a_1}{\alpha^2} + \cdots + \frac{a_{r-1}}{\alpha^r} + \mathrm{O}(\alpha^{-r-1}), \tag{6.10}$$

where

$$a_k = \int_0^\infty \rho(t) t^k \mathrm{d}t. \tag{6.11}$$

It remains to evaluate the a_k. We have

$$\sum_{k \geqslant 0} a_k \frac{(-x)^k}{k!} = \int_0^\infty \rho(t) \mathrm{e}^{-xt} \mathrm{d}t = e_1(x) = \mathrm{e}^{-E(x) - \ln x} \tag{6.12}$$

by Theorem 5.2; and it is well known that

$$-E(x)-\ln x=\gamma+\sum_{k\geqslant 1}(-x)^k/k\cdot k!. \tag{6.13}$$

(See, for example, [8, exercise 5.2.2–43].) This combines with (6.12) and (6.6) to prove that $a_k=e^\gamma c_k$. □

Incidentally, a similar method can be used to show that $\int_1^\infty(\rho_2(t)-e^\gamma/t)\mathrm{d}t=e^\gamma-1$.

The coefficients c_k have the curious property that

$$\rho_2(\alpha)=A\left(\frac{c_0}{\alpha+1}+\frac{2c_1}{(\alpha+1)^2}+\cdots+\frac{rc_{r-1}}{(\alpha+1)^r}\right)+O(\alpha^{-r-1}) \tag{6.14}$$

is also an asymptotic expansion of ρ_2, but not as accurate when truncated. Another series,

$$\rho_2(\alpha)=A\left(\frac{c_0}{\alpha-1}+\frac{c_1-c_0}{2(\alpha-1)^2}+\frac{c_2-c_1+c_0}{3(\alpha-1)^3}+\cdots\right)+O(\alpha^{-r-1})$$

is, in turn, more accurate than (6.4). These series are obtainable from one another using the relation $\rho_2(\alpha)=-(\alpha+1)\rho_2'(\alpha+1)+\rho_1(\alpha)$.

For $k\geqslant 3$, we shall content ourselves with establishing the leading term in the asymptotic expansion of ρ_k, namely

$$\rho_k(\alpha)=\frac{A(\ln\alpha)^{k-2}}{(k-2)!\,\alpha}+O\left(\frac{(\ln\alpha)^{k-3}}{\alpha}\right),\quad \text{for } k\geqslant 3. \tag{6.15}$$

[Appendix B contains an asymptotic expansion of ρ_3.] Consider first

$$S_2(\alpha)=\int_1^\alpha\left(\frac{A}{t}+O\left(\frac{1}{t^2}\right)\right)\frac{\mathrm{d}t}{\alpha+1-t} \tag{6.16}$$

and note that

$$\int_1^\alpha\frac{\mathrm{d}t}{t(\alpha+1-t)}=\frac{1}{\alpha+1}\left(\int_1^\alpha\frac{\mathrm{d}t}{t}+\int_1^\alpha\frac{\mathrm{d}t}{\alpha+1-t}\right)=\frac{2\ln\alpha}{\alpha+1},$$

$$\int_1^\alpha\frac{\mathrm{d}t}{t^2(\alpha+1-t)}=\frac{1}{\alpha+1}\int_1^\alpha\frac{\mathrm{d}t}{t^2}+\frac{1}{\alpha+1}\int_1^\alpha\frac{\mathrm{d}t}{t(\alpha+1-t)}$$

$$=\frac{1}{\alpha+1}\left(1-\frac{1}{\alpha}\right)+\frac{2\ln\alpha}{(\alpha+1)^2}=O(\alpha^{-1}). \tag{6.17}$$

Hence $S_2(\alpha)=2A\alpha^{-1}\ln\alpha+O(\alpha^{-1})$, and $\rho_3(\alpha)=A\alpha^{-1}\ln\alpha+O(\alpha^{-1})$ by (5.18). In order to use this approach for larger k, we note that, when $k\geqslant 1$,

$$\int_1^\alpha\frac{(\ln t)^k\mathrm{d}t}{t(\alpha+1-t)}=\frac{1}{\alpha+1}\int_1^\alpha\frac{(\ln t)^k\mathrm{d}t}{t}+\frac{1}{\alpha+1}\int_1^\alpha\frac{(\ln t)^k\mathrm{d}t}{\alpha+1-t}$$

$$=\frac{1}{(\alpha+1)}\frac{(\ln\alpha)^{k+1}}{(k+1)}+\frac{k}{\alpha+1}\int_1^\alpha\frac{(\ln t)^{k-1}\ln(\alpha+1-t)\mathrm{d}t}{t}$$

*Ed: Eq # (6.16)?

$$=\frac{1}{(\alpha+1)}\frac{(\ln\alpha)^{k+1}}{(k+1)}+\frac{\ln(\alpha+1)}{\alpha+1}(\ln\alpha)^k$$

$$+\frac{k}{\alpha+1}\int_1^\alpha\frac{(\ln t)^{k-1}\ln\left(1-\frac{t}{\alpha+1}\right)\mathrm{d}t}{t}.$$

Now $\ln(1-x)=-xf(x)$, where f is a function satisfying

$$1<f\left(\frac{t}{\alpha+1}\right)\leqslant f\left(\frac{\alpha}{\alpha+1}\right)=\frac{\alpha+1}{\alpha}\ln(\alpha+1) \tag{6.18}$$

when $1\leqslant t\leqslant\alpha$, hence

$$\int_1^\alpha\frac{(\ln t)^{k-1}}{t}\ln\left(1-\frac{t}{\alpha+1}\right)\mathrm{d}t=\frac{1}{\alpha+1}\int_1^\alpha(\ln t)^{k-1}\mathrm{O}(\ln\alpha)\mathrm{d}t$$
$$=\mathrm{O}(\ln\alpha)^k. \tag{6.19}$$

We have proved that

$$\int_1^\alpha\frac{(\ln t)^k\mathrm{d}t}{t(\alpha+1-t)}=\frac{k+2}{k+1}\frac{(\ln\alpha)^{k+1}}{\alpha}+\mathrm{O}\left(\frac{(\ln\alpha)^k}{\alpha}\right), \tag{6.20}$$

for all $k\geqslant 0$. Using (5.18), formula (6.15) now follows by induction, together with

$$S_k(\alpha)=\frac{\mathrm{A}k(\ln\alpha)^{k-1}}{(k-1)!\,\alpha}+\mathrm{O}\left(\frac{(\ln\alpha)^{k-2}}{\alpha}\right). \tag{6.21}$$

7. Application to factoring

The distributions $F_k(x)=\rho_k(1/x)$ can be used to estimate the running time of various algorithms for factorization. For example, Pollard's important new Monte Carlo method [11] takes about $\sqrt{n_2}$ steps, where n_2 is the second-largest prime factor of n, so we can use a table of F_2 to state that Pollard's method will complete the factorization in $\mathrm{O}(n^{0.106})$ steps at most, about half of the time.

For the simple algorithm of Section 1, we need to analyze the distribution of $\max(n_2,\sqrt{n_1})$, and this does not appear to be expressible directly as an algebraic function of the F_k. However, we can readily carry out the analysis by using the techniques above. Let $G(x)$ be the limiting probability that $\max(n_2,\sqrt{n_1})\leqslant N^x$, when n is a random integer between 1 and N. Then $G(x)=F_1(x)+G_1(x)=F_2(x)-G_2(x)$, where $G_1(x)$ is the probability that $N^x\leqslant n_1\leqslant N^{2x}$ and $n_2\leqslant N^x$, and $G_2(x)$ is the probability that $n_1>N^{2x}$ and $n_2\leqslant N^x$. Arguing as above, we find

$$G_1(x)=\int_x^{2x}\frac{\mathrm{d}t}{t}F_1\left(\frac{x}{1-t}\right)=\int_x^{2x}\frac{\mathrm{d}t}{t}\rho\left(\frac{1-t}{x}\right), \tag{7.1}$$

$$G_1\left(\frac{1}{\alpha}\right) = \int_{1/\alpha}^{2/\alpha} \rho((1-t)\alpha)\,\frac{\mathrm{d}t}{t} = \int_{\alpha-1}^{\alpha} \frac{\rho(u-1)\mathrm{d}u}{\alpha+1-u}\,, \tag{7.2}$$

$$G_2(x) = \int_{2x}^{1} \frac{\mathrm{d}t}{t}\,F_1\left(\frac{x}{1-t}\right) = \int_{2x}^{1} \frac{\mathrm{d}t}{t}\,\rho\left(\frac{1-t}{x}\right), \tag{7.3}$$

$$G_2\left(\frac{1}{\alpha}\right) = \int_{2/\alpha}^{1} \rho((1-t)\alpha)\,\frac{\mathrm{d}t}{t} = \int_{1}^{\alpha-1} \frac{\rho(u-1)\mathrm{d}u}{\alpha+1-u}\,. \tag{7.4}$$

(Note that $G_1(1/\alpha)+G_2(1/\alpha)=S_1(\alpha)=F_2(1/\alpha)-F_1(1/\alpha)$, in agreement with Lemma 5.1.) It is clear from our asymptotic results that $G_1(1/\alpha)$ decreases exponentially for large α, hence it is numerically better to use the formula $G(x)=F_1(x)+G_1(x)$ than to use $F_2(x)-G_2(x)$; furthermore the integration is over a limited range. On the other hand for $2 \leqslant \alpha \leqslant 3$ it is most convenient to use G_2 since $G_2(1/\alpha)=\ln(\alpha/2)$ in this range.

8. Remarks about the model

The probability considerations above are for random n between 1 and N, and for relations such as $n_k \leqslant N^x$; but from an intuitive standpoint we might rather ask for the probability of a relation such as $n_k \leqslant n^x$, without considering N. Actually it is easy to convert from one model to the other, since most numbers between 1 and N are large.

More precisely, consider how many numbers n between $\frac{1}{2}N$ and N have $n_k \leqslant N^x$; this is $P_k(x,N)-P_k(x,(1/2)N)=(1/2)N\cdot F_k(x)+\mathrm{O}(N/\log N)$, since $P_k(x,N)=N\cdot F_k(x)+\mathrm{O}(N/\log N)$. Furthermore, consider how many of these n have $n^x < n_k \leqslant N^x$: The latter relation implies $N^x \geqslant n_k > (\frac{1}{2}N)^x = N^{x-\log 2/\log N}$, and $F_k(x-\log 2/\log N)=F_k(x)+\mathrm{O}(1/\log N)$, since F_k is differentiable; so the number of such n is at most $P_k(x,N)-P_k(x-\log 2/\log N, N)=\mathrm{O}(N/\log N)$. (The constant implied by the O in (4.7) will be independent of x in a bounded region about x.)

We have shown that $F_k(x)+\mathrm{O}(1/\log N)$ of all n between $\frac{1}{2}N$ and N satisfy $n_k \leqslant n^x$. Therefore if $Q_k(x,N)$ denotes the total number of $n \leqslant N$ such that $n_k \leqslant n^x$, we have

$$\begin{aligned} Q_k(x,N) &= \sum_{1\leqslant j\leqslant \log_2 \log N} \frac{1}{2^j}\,N\left(F_k(x)+\mathrm{O}\left(\frac{1}{\log(N/2^j)}\right)\right)+\mathrm{O}\left(\frac{N}{\log N}\right) \\ &= NF_k(x)+\mathrm{O}\left(\frac{N}{\log N}\right), \end{aligned} \tag{8.1}$$

by dividing the range $N/\log N \leqslant n \leqslant N$ into $\log_2 \log N$ parts.

It is customary to define the "probability" of a statement $S(n)$ about the positive integer n by the formula

$$\Pr(S(n)) = \lim_{N\to\infty} \frac{1}{N} \text{ (number of } n \leqslant N \text{ such that } S(n) \text{ is true)}, \tag{8.2}$$

when this limit exists. Thus, we can state well-known facts such as the following: $\Pr(n \text{ is even}) = 1/2$; $\Pr(n \text{ is prime}) = 0$; $\Pr(n \text{ is squarefree}) = 6\pi^2$. Eq. 8.1 now yields another result of this type:

$$\Pr(n_k \leqslant n^x) = F_k(x), \tag{8.3}$$

for all fixed x.

Another important observation should also be made about the theoretical model we have used to study the factorization algorithm in this paper. We have stated our results in terms of the probability that the running time is $\leqslant N^x$ (or, if we prefer, n^x); this contrasts with the customary approach to the study of average running time, which derives mean values and the standard deviation. The reason for abandoning the traditional approach is that the mean and standard deviation are particularly uninformative for this algorithm. This phenomenon is apparent when we consider that the mean running time over all $n \leqslant N$ will be relatively near the worst case $n^{0.5}$, but in more than 70 per cent of all cases the actual running time will be less than $n^{0.4}$.

In order to understand this rather anomalous situation more fully, let us calculate the asymptotic mean and standard deviation of the largest prime factor n_1, when all integers $1 \leqslant n \leqslant N$ are considered equally likely. Let $\Phi(t)$ be the probability that $n_1 \leqslant t$, when n is in this range. Then the derivation of (4.13) allows us to conclude that

$$\Phi(t) = 1 + \ln\ln t - \ln\ln N + \frac{1}{\ln N}\int_1^{N/t} \frac{\{u\}\mathrm{d}u}{u^2} + \mathrm{O}\left(\frac{1}{\log N}\right)^2, \tag{8.4}$$

for $\sqrt{N} \leqslant t \leqslant N$.

We shall now calculate the asymptotic behavior of the k^{th} moment of this distribution, namely the asymptotic expected value of n_1^k. [Incidentally, our derivation provides a good example of the use of Stieltjes integration.] The k^{th} moment is

$$\mathrm{E}(n_1^k) = \int_1^N t^k \mathrm{d}\Phi(t), \tag{8.5}$$

and since the integral from $1 + \sqrt{N}$ is $\mathrm{O}(N^{k/2}\int_1^{\sqrt{N}} \mathrm{d}\Phi(t)) = \mathrm{O}(N^{k/2})$ it can safely be ignored. We are left with

$$\int_{\sqrt{N}}^N t^k \mathrm{d}\left(1 + \ln\ln t - \ln\ln N + \frac{1}{\ln N}\int_1^{N/t} \frac{\{u\}\mathrm{d}u}{u^2} + \mathrm{O}\left(\frac{1}{\log N}\right)^2\right)$$

$$= \int_{\sqrt{N}}^N t^k\, \mathrm{d}(\ln\ln t) + \frac{1}{\ln N}\int_{\sqrt{N}}^1 \left(\frac{N}{v}\right)^k \mathrm{d}\int_1^v \frac{\{u\}\,du}{u^2} + \mathrm{O}\left(\frac{N^k}{(\log N)^2}\right), \tag{8.6}$$

by replacing t by N/v in the second integral. [The O estimate here is justified by the following general lemma: Let $\int_a^b f(t)\mathrm{d}g(t)$ and $\int_a^b f(t)\mathrm{d}h(t)$ exist, where $h(t) = \mathrm{O}(g(t))$, and where both f and g are positive monotone functions on $[a, b]$. Then it is easy to see that

$$\int_a^b f(t)\mathrm{d}\mathrm{O}(g(t)) = \mathrm{O}(f(a)g(a)) + \mathrm{O}(f(b)g(b)) + \mathrm{O}\left(\int_a^b f(t)\mathrm{d}g(t)\right),$$

if we integrate by parts twice.] The first integral in (8.6) is

$$\int_{\sqrt{N}}^{N} \frac{t^{k-1}\mathrm{d}t}{\ln t} = N^k \int_1^{\sqrt{N}} \frac{\mathrm{d}v}{v^{k+1}(\ln N - \ln v)} = \frac{N^k}{\ln N}\left(\int_1^{\sqrt{N}} \frac{\mathrm{d}v}{v^{k+1}} + \int_1^{\sqrt{N}} \frac{\ln v\,\mathrm{d}v}{v^{k+1}(\ln N - \ln v)}\right)$$
$$= \frac{N^k}{k\ln N} + \mathrm{O}\left(\frac{N^k}{(\log N)^2}\right).$$

The second integral is $-N^k/\ln N$ times the integral $\int_1^{\sqrt{N}}\{v\}\mathrm{d}v\,/\,v^{k+2}$, which is within $\mathrm{O}(N^{-(k+1)/2})$ of

$$\int_1^{\infty} \frac{\{v\}\mathrm{d}v}{v^{k+2}} = \sum_{j\geq 1} \int_j^{j+1} \frac{(v-j)\mathrm{d}v}{v^{k+2}}$$
$$= \sum_{j\geq 1}\left(\frac{1}{k}\left(\frac{1}{j^k} - \frac{1}{(j+1)^k}\right) - \frac{j}{k+1}\left(\frac{1}{j^{k+1}} - \frac{1}{(j+1)^{k+1}}\right)\right)$$
$$= \sum_{j\geq 1}\left(\frac{1}{k(k+1)}\left(\frac{1}{j^k} - \frac{1}{(j+1)^k}\right) - \frac{1}{k+1}\,\frac{1}{(j+1)^{k+1}}\right)$$
$$= \frac{1}{k(k+1)} - \frac{1}{k+1}(\zeta(k+1) - 1) = \frac{1}{k} - \frac{\zeta(k+1)}{k+1}.$$

Thus we have shown that

$$\mathrm{E}(n_1^k) = \frac{\zeta(k+1)}{k+1}\,\frac{N^k}{\ln N} + \mathrm{O}\left(\frac{N^k}{(\log N)^2}\right). \tag{8.7}$$

It follows that the mean value of n_1 is asymptotically $(\pi^2/12)N/\ln N$, and the standard deviation is $(\zeta(3)/3)^{1/2}N/\sqrt{\ln N}$, to within a factor of $1 + \mathrm{O}(1/\log N)$. In particular, the ratio

$$\frac{\text{standard deviation}}{\text{mean}} \to \infty \tag{8.8}$$

as $N \to \infty$; this result demonstrates the unsuitability of a traditional "mean and variance" approach to the analysis of such algorithms.

9. Numerical results

The differential-difference equations for ρ_k are conveniently suited to numerical integration. For example, given internal arrays containing $\rho_1(m + k/n)$, $\rho_2(m + k/n)$, and $\rho_3(m + k/n)$ for $0 \leq k \leq n + t$, where m is some fixed integer and $\delta = 1/n$ is the step size and t depends on the method of integration, one pass over these arrays serves to increase m by 1. When m reaches a suitably large value, the asymptotic formulas derived above provide an excellent check on the accuracy of the calculations. Another excellent check comes from the formula

$$e^{\gamma} = \int_0^{\infty} \rho(t)\mathrm{d}t = \rho(1) + 2\rho(2) + 3\rho(3) + \cdots ; \tag{9.1}$$

cf. (6.2), (6.5), and (6.11). (Incidentally, identity (9.1) appears to be new; it was discovered empirically, after noticing that the results of numerical integration seemed to resemble a "familiar" constant. This particular constant came as a surprise, since e^{γ} usually occurs only in connection with infinite products. After the proof of (9.1) was found, Theorem 5.2 followed rather quickly. Thus, numerical results indeed suggest theorems.)

The following table gives representative values of ρ_1, ρ_2, ρ_3 and G to $12D$:

Table 1

α	$\rho_1(\alpha)$	$\rho_2(\alpha)$	$\rho_3(\alpha)$	$G(1/\alpha)$
1.0	1.000000 000000	1.000000 000000	1.000000 000000	1.000000 000000
1.5	0.594534 891892	1.000000 000000	1.000000 000000	1.000000 000000
2.0	0.306852 819440	1.000000 000000	1.000000 000000	1.000000 000000
2.5	0.130319 561832	0.953389 706294	1.000000 000000	0.730246 154979
3.0	0.048608 388291	0.852779 323041	1.000000 000000	0.447314 214932
3.5	0.016229 593243	0.733481 165219	0.997526 273042	0.223819 493955
4.0	0.004910 925648	0.623681 059959	0.985113 653272	0.096399 005935
4.5	0.001370 117741	0.533652 572034	0.960975 011157	0.036573 065077
5.0	0.000354 724700	0.463222 186987	0.927859 653628	0.012413 482748
6.0	0.000019 649696	0.365217 751694	0.851107 195638	0.001092 266742
7.0	0.000000 874567	0.301786 010308	0.777229 329492	0.000071 391673
8.0	0.000000 032321	0.257435 710831	0.712844 794121	0.000003 662651
9.0	0.000000 001016	0.224592 162720	0.657959 581954	0.000000 153284
10.0	0.000000 000028	0.199248 208994	0.611115 997540	0.000000 005383
12.0	0.000000 000000	0.162638 856635	0.535865 613616	0.000000 000004
14.0	0.000000 000000	0.137437 368144	0.478221 749442	0.000000 000000
16.0	0.000000 000000	0.119016 453035	0.432642 865532	0.000000 000000
18.0	0.000000 000000	0.104958 753569	0.395653 753569	0.000000 000000
20.0	0.000000 000000	0.093875 845625	0.364991 546696	0.000000 000000
25.0	0.000000 000000	0.074277 803044	0.307069 057805	0.000000 000000
30.0	0.000000 000000	0.061453 736517	0.266170 912880	0.000000 000000
40.0	0.000000 000000	0.045683 813582	0.211838 770538	0.000000 000000
50.0	0.000000 000000	0.036356 095670	0.177085 969207	0.000000 000000
60.0	0.000000 000000	0.030192 055732	0.152778 425203	0.000000 000000

(In 1930, Dickman published $8D$ values of $\rho_1(\alpha)$ for integer $\alpha \leqslant 8$; his figures were correct except that $\rho_1(7)$ was given as "0.0000 0088".)

Fig. 1 shows these distributions graphically, and illustrates the fact that $F_1'(0) = G'(0) = F_2'(\frac{1}{2}) = F_3'(\frac{1}{3}) = 0$, $F_2'(0) = A$, $G'(\frac{1}{2}) = 2$, $F_1'(1) = 1$, $F_3'(0) = \infty$. Although the graphs of F_1, F_2, and F_3 are qualitatively different, the graphs of F_k for $k \geqslant 4$ will resemble that of F_3 (but they will rise ever more steeply).

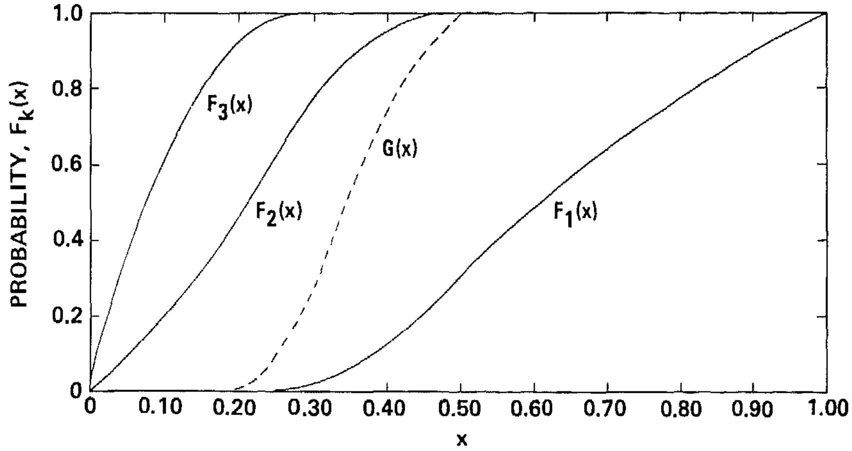

Fig. 1. Distributions of the three largest prime factors of a random integer and the distribution of the simple factorization time.

Table 2 shows percentage points of the distributions F_1, F_2, F_3; for example, the probability is only 10 percent that $n_3 > n^{0.18616}$.

Table 2

p	$F_1^{-1}(p)$	$F_2^{-1}(p)$	$F_3^{-1}(p)$
0.01	0.26974	0.00558	0.00068
0.02	0.29341	0.01110	0.00149
0.03	0.31004	0.01656	0.00239
0.04	0.32341	0.02196	0.00334
0.05	0.33483	0.02730	0.00435
0.10	0.37851	0.05308	0.00995
0.15	0.41288	0.07741	0.01629
0.20	0.44304	0.10033	0.02327
0.25	0.47068	0.12191	0.03079
0.30	0.49656	0.14216	0.03882
0.40	0.54881	0.17892	0.05636
0.50	0.60653	0.21172	0.07584
0.60	0.67032	0.24267	0.09745
0.70	0.74082	0.27437	0.12165
0.75	0.77880	0.29153	0.13506
0.80	0.81873	0.31035	0.14972
0.85	0.86071	0.33201	0.16627
0.90	0.90484	0.35899	0.18616
0.95	0.95123	0.39672	0.21377
0.96	0.96079	0.40681	0.22141
0.97	0.97045	0.41850	0.23054
0.98	0.98020	0.43268	0.24224
0.99	0.99005	0.45169	0.25954
1.00	1.00000	0.50000	0.33333

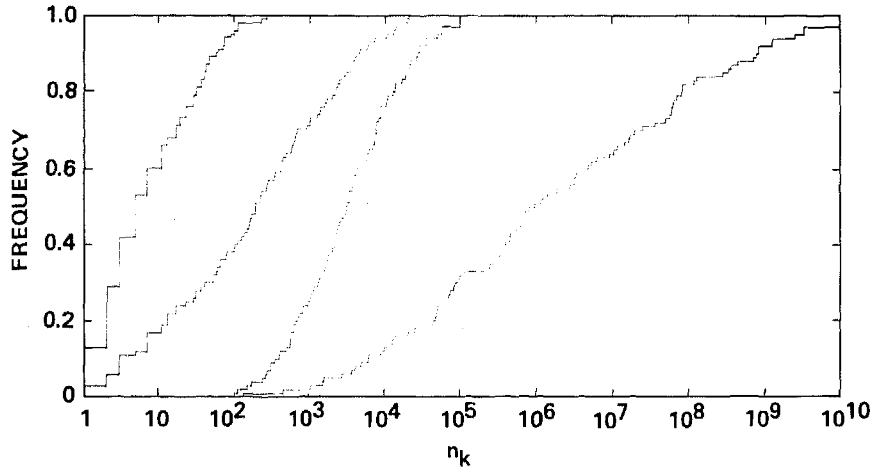

Fig. 2. Empirical distribution functions corresponding to Fig. 1, based on the factors of the largest 100 10-digit numbers.

Empirical confirmation of the theory is illustrated in Fig. 2, which shows exact empirical distribution functions corresponding to Fig. 1 for the 100 numbers $n = 10^{10}–m$, $1 \leq m \leq 100$. As expected, the deviation from $F_k(x)$ is most pronounced for $k = 1$ and $x > \frac{1}{2}$, but the deviations are not severe. This set of numbers contains three primes (10^{10}–33, 10^{10}–57, 10^{10}–71), and ten products of two primes. The smallest values of n_1 occurred for $10^{10} - 100 = 137 \cdot 101 \cdot 73 \cdot 11 \cdot 5^2 \cdot 3^2 \cdot 2^2$, $10^{10}–64 = 463 \cdot 431 \cdot 29 \cdot 3^3 \cdot 2^6$; the largest values of n_2 occurred for $10^{10}–69 = 456767 \cdot 21893$, $10^{10}–22 = 85021 \cdot 19603 \cdot 3 \cdot 2$; the largest values of n_3 occurred for $10^{10}–51 = 88301 \cdot 421 \cdot 269$, $10^{10}–73 = 13879 \cdot 359 \cdot 223 \cdot 3^2$. The smallest values of $\max(\sqrt{n_1}, n_2)$ occurred for $10^{10}–100 = 137 \cdot 101 \cdot 73 \cdot 11 \cdot 5^2 \cdot 3^2 \cdot 2^2$, $10^{10}–25 = 2857 \cdot 113 \cdot 59 \cdot 7 \cdot 5^2 \cdot 3$ (so these would be the easiest numbers in the given range to factor by the simple algorithm); the smallest values of n_1 for which $\sqrt{n_1} > n_2$ occurred for $10^{10}–66 = 59417 \cdot 103 \cdot 43 \cdot 19 \cdot 2$, $10^{10}–68 = 77201 \cdot 53 \cdot 47 \cdot 13 \cdot 2^2$.

In Dickman's original paper he calculated the "average" value of x such that $n_1 = n^x$, namely the expected value of $\log n_1 / \log n$. This equals

$$D_1 = \int_0^1 x \, \mathrm{d}F_1(x) = -\int_1^\infty \rho'(t)\mathrm{d}t/t = \int_1^\infty \rho(t-1)\mathrm{d}t/t^2 \qquad (9.2)$$

and by (5.14) we also have

$$\int_1^\infty \rho(t-1)\mathrm{d}t/t^2 = -S_1(\infty, -1) = \int_1^\infty \rho(t-1)\mathrm{d}t/(t+1). \qquad (9.3)$$

In a similar way we can determine the expected value of $\log n_k / \log n$, a number which can be expressed in several ways, namely

$$D_k = \int_0^1 x\, \mathrm{d}F_k(x) = \int_1^\infty (\rho_k(t-1) - \rho_{k-1}(t-1))\mathrm{d}t/t^2 = 1 - \int_1^\infty \rho_k(t)\mathrm{d}t/t^2$$

$$= \int_1^\infty (\rho_k(t-1) - 2\rho_{k-1}(t-1) + \rho_{k-2}(t-1))\mathrm{d}t/(t+1). \quad (9.4)$$

Numerical evaluation (using the asymptotic formulas for ρ_2 and ρ_3) gives

$$D_1 = 0.62432\ 99885; \quad (9.5)$$

$$D_2 = 0.20958\ 08743; \quad (9.6)$$

$$D_3 = 0.08831\ 60989. \quad (9.7)$$

(Dickman's value for D_1 was 0.624329998. Note that D_2 is not equal to $D_1(1-D_1)$, although n_2 is the largest prime factor of n/n_1.)

The average value of a logarithm may seem at first to be of limited practical interest, by comparison with the median and other percentiles; however, we can interpret it meaningfully by saying that $D_k m$ is the *asymptotic average number of digits* in the k^{th} largest prime factor of an m-digit number. Dickman's constant D_1 arises also in an unexpected way in connection with our simple factoring algorithm: The probability that $n_2 < \sqrt{n_1}$, namely the probability that the algorithm needs to divide by all numbers up to $\sqrt{n_1}$, is

$$\int_0^1 \frac{\mathrm{d}t}{t} F_1\left(\frac{t}{2(1-t)}\right) = \int_0^1 \frac{\mathrm{d}t}{t} \rho\left(\frac{2(1-t)}{t}\right) = \int_1^\infty \rho(u-1)\mathrm{d}u/(u+1) \quad (9.8)$$

by substituting $u = 2/t - 1$. So this probability equals D_1! In the empirical tests which led to Fig. 2, exactly 61 of the 100 numbers had $n_2 < \sqrt{n_1}$.

10. Relation to permutations

The numerical value of D_1 in (9.5) leads again to a feeling of *déjà vu*; and sure enough Dickman's constant turns out to be the same as "Golomb's constant", which has been evaluated to 53 places in [9]. Golomb's constant λ is defined to be $\lim_{n\to\infty} l_n/n$, where l_n is the average length of the longest cycle in a random n-permutation. In Golomb's original analysis [6] of this combinatorial problem (which is not *obviously* related to prime factors at all!), he independently defined a function essentially identical to $\rho(\alpha)$, and he computed $\lambda = \int_1^\infty \rho(t-1)\mathrm{d}t/t^2$ numerically. Another expression $\lambda = \int_1^\infty \exp(-x - E(x))\mathrm{d}x$ was found later by Shepp and Lloyd [13].

In Table 1 of their paper, Shepp and Lloyd also list the limiting values $l^{(k)}/n \to \int_0^\infty E(t)^{k-1}\exp(-t - E(t))\mathrm{d}t/(k-1)!$ for the average length of the k^{th} longest cycle; and this agrees numerically with D_k for $1 \leqslant k \leqslant 3$. In fact, the Shepp–Lloyd formula yields D_k for all k, since

$$\begin{aligned}\int_0^\infty \frac{E(t)^{k-1}}{(k-1)!}\exp(-t-E(t))\mathrm{d}t &= \int_0^\infty t\,\mathrm{e}^{-t}(e_k(t)-e_{k-1}(t))\mathrm{d}t\\ &= \int_0^\infty t\,\mathrm{e}^{-t}\int_0^\infty (\rho_k(u)-\rho_{k-1}(u))\mathrm{e}^{-tu}\mathrm{d}u\;\mathrm{d}t\\ &= \int_1^\infty (\rho_k(u-1)-\rho_{k-1}(u-1))\int_0^\infty t\,\mathrm{e}^{-t(u)}\mathrm{d}t\,\mathrm{d}u\\ &= \int_1^\infty (\rho_k(u-1)-\rho_{k-1}(u-1))\mathrm{d}u/u^2. \qquad (10.1)\end{aligned}$$

Therefore, if we are factoring a random m-digit number, the distribution of the number of digits in its prime factors is *approximately the same as the distribution of the cycle lengths in a random permutation* on m elements! (Note that there are approximately $\ln m$ factors, and approximately $\ln m$ cycles.)

There is a fairly simple explanation for the fact that $\rho_k(\alpha)$ turns up in the study of cycles in permutations. Let $Q_k(n,r)$ be the number of permutations on n objects having less than k cycles of length exceeding r. Then, by considering the permutations on $n+1$ elements $\{0,1,\ldots,n\}$ and considering the $n!/(n-m)!$ possible cycles in which 0 appears with m different elements, we have

$$Q_k(n+1,r) = \sum_{0\leq m<r}\frac{n!}{(n-m)!}Q_k(n-m,r) + \sum_{r\leq m\leq n}\frac{n!}{(n-m)!}Q_{k-1}(n-m,r). \qquad (10.2)$$

Therefore if $q_k(n,r) = Q_k(n,r)/n!$ is the probability that the k^{th} largest cycle has length $\leq r$, we have

$$(n+1)q_k(n+1,r) = \sum_{0\leq m<r} q_k(n-m,r) + \sum_{r\leq m\leq n} q_{k-1}(n-m,r); \qquad (10.3)$$

replacing n by $n-1$ yields

$$nq_k(n,r) = \sum_{0\leq m<r} q_k(n-1-m,r) + \sum_{r\leq m\leq n} q_{k-1}(n-1-m,r). \qquad (10.4)$$

Subtracting these two equations, we have

$$(n+1)(q_k(n+1,r)-q_k(n,r)) = q_{k-1}(n-r,r)-q_k(n-r,r), \qquad (10.5)$$

and this is analogous to the differential equation

$$\alpha\rho_k'(\alpha) = \rho_{k-1}(\alpha-1)-\rho_k(\alpha-1). \qquad (10.6)$$

The connection between the two problems is completed by showing that $q_k(n,r) = \rho_k(n/r) + \mathrm{O}(1/r)$.

A similar distribution is obtained for the degrees of the factors of a random polynomial of degree n, over a finite field: The average degree of the k^{th} "largest" irreducible factor will tend to be approximately $D_k n$.

Let us close by stating an open problem: Are the functions ρ_k algebraically independent? They are linearly independent, because of (5.5).

Acknowledgments

We wish to thank N.G. de Bruijn, Roger Eggleton, Karl K. Norton, John Pollard, and Marvin Wunderlich for correspondence relating to early drafts of this paper. A letter received from Nathan. J. Fine in June, 1965, regarding Golomb's problem, was also very helpful for this research.

Appendix A
The number of prime factors

Following the notation of Hardy and Wright [7], let $\omega(n)$ be the number of distinct prime factors of n, and let $\Omega(n)$ be the total number of prime factors including multiplicity. Thus, $\Omega(n)$ is the quantity T in the analysis of the algorithm above. Clearly $1 \leqslant \Omega(n) \leqslant \log_2 n$, and both of these limits are obtained for infinitely many n; similarly $\omega(n)$ can get as large as $\ln n/\ln\ln n$. On the other hand these extreme values are relatively rare, and the number of factors is usually near $\ln\ln n$.

Erdös and Kac [5] proved that the number of n in the range $1 \leqslant n \leqslant N$ such that $\omega(n) < \ln\ln N + c\sqrt{\ln\ln N}$ is

$$\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{c} \mathrm{e}^{-t^2/2}\mathrm{d}t\right) N + \mathrm{o}(N); \tag{A.1}$$

hence, for example, the probability that $|\omega(n) - \ln\ln N| < c\sqrt{\ln\ln N}$ for fixed $c > 0$ approaches the limiting value

$$\frac{1}{\sqrt{2\pi}}\int_{-c}^{c} \mathrm{e}^{-t^2/2}\mathrm{d}t. \tag{A.2}$$

We might say that $\omega(n)$ behaves essentially like a normally distributed random variable with mean and variance $\ln\ln n$, where n is large.

Erdös and Kac remarked that their methods, which were based on the idea that residues modulo distinct primes are independent, could be extended to the case of prime factors with multiplicities included, but they did not state what the resulting theorem would be. Fortunately it is easy to deduce the asymptotic behavior of $\Omega(n)$ from that of $\omega(n)$, using a method like that in [6]. Let $k(N)$ be the number of n in $1 \leqslant n \leqslant N$ such that

$$\omega(n) < \ln\ln N + c\sqrt{\ln\ln N} \tag{A.3}$$

and let $K(N)$ be the number such that

$$\Omega(n) < \ln\ln N + c\sqrt{\ln\ln N} + \ln\ln\ln N. \tag{A.4}$$

Then $|k(N) - K(N)|$ is at most the number of n which satisfy (A.3) but not (A.4), or (A.4) but not (A.3), and both of these quantities are o(N): If n satisfies (A.3) but not (A.4), we have $\Omega(n) - \omega(n) > \ln\ln\ln N$; and the number of such n is $O(N/\ln\ln\ln N)$, because

$$\sum_{1 \leq n \leq N} (\Omega(n) - \omega(n)) = O(N) \tag{A.5}$$

by [7, Theorem 430]. If n satisfies (A.4) but not (A.3), then

$$\ln\ln N + c\sqrt{\ln\ln N} \leq \omega(n) < \ln\ln N + \left(c + \frac{\ln\ln\ln N}{\sqrt{\ln\ln N}}\right)\sqrt{\ln\ln N},$$

and this is o(N) by the theorem of Erdös and Kac.

We have proved that the number of n in the rage $1 \leq n \leq N$ such that $\Omega(n) < \ln\ln N + c\sqrt{\ln\ln N}$ is asymptotically given by the normal distribution (A.1). But this estimate is insensitive to O(1) terms, so the "average order" [7, Theorem 430] is also relevant:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1 \leq n \leq N} (\omega(n) - \ln\ln N)$$

$$= \gamma + \sum_{p \text{ prime}} \left(\log\left(1 - \frac{1}{p}\right) + \frac{1}{p}\right) \approx 0.26149\ 72128\ 47643; \tag{A.6}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1 \leq n \leq N} (\Omega(n) - \ln\ln N)$$

$$= \gamma + \sum_{p \text{ prime}} \left(\log\left(1 - \frac{1}{p}\right) + \frac{1}{p-1}\right) \approx 1.03465\ 38818\ 97438. \tag{A.7}$$

These sums may be evaluated to high precision using the formula

$$\sum_{p \text{ prime}} \frac{1}{p^s} = \sum_{n \geq 1} \frac{\mu(n)}{n} \ln \zeta(ns), \quad \text{for } s > 1. \tag{A.8}$$

Let $S = \{10^{10} - m \mid 1 \leq m \leq 100\}$ be the numbers used to construct Fig. 2 above. For $n \in S$ we have $\ln\ln n \approx 3.1366$, and the following table shows the actual distribution of $\omega(n)$ and $\Omega(n)$.

Table 3

k	1	2	3	4	5	6	7	8	9	10	11	12
$\Vert\{n \in S \mid \omega(n) = k\}\Vert$	3	14	36	29	14	3	1	0	0	0	0	0
$\Vert\{n \in S \mid \Omega(n) = k\}\Vert$	3	10	27	23	15	11	5	3	1	1	0	1

The respective mean values are 3.50 and 4.27. The number of square-free n (those with $\omega(n) = \Omega(n)$) was 61, compared to the expected value $600/\pi^2 = 60.793$.

Appendix B
An asymptotic formula for ρ_3

In this appendix we shall sketch the derivation of an asymptotic expression for $\rho_3(\alpha)$ as $\alpha \to \infty$. Our starting point is the formula

$$\begin{aligned} S_2(\alpha) &= \int_0^{\alpha-1} \frac{\rho_2(t)\mathrm{d}t}{\alpha - t} \\ &= \sum_{0 \leqslant k \leqslant r} \frac{1}{\alpha^{k+1}} \int_0^{\alpha-1} \rho_2(t) t^k \, \mathrm{d}t + \frac{1}{\alpha^{r+1}} \int_0^{\alpha-1} \frac{\rho_2(t) t^{r+1} \mathrm{d}t}{\alpha - t} ; \qquad \text{(B.1)} \end{aligned}$$

we replace the final term by its asymptotic value

$$\frac{A}{\alpha^{r+1}} \int_0^{\alpha-1} \frac{(c_0 t^r + c_1 t^{r-1} + \cdots + c_{r-1} t)\mathrm{d}t}{\alpha - t} + \mathrm{O}\left(\frac{1}{\alpha^{r+1}} \int_0^{\alpha-1} \frac{\mathrm{d}t}{\alpha - t}\right), \qquad \text{(B.2)}$$

so that the remainder is $\mathrm{O}(\alpha^{-r-1} \log \alpha)$. The main integral in (B.2) is a linear combination of

$$\int_0^{\alpha-1} \frac{t^k \mathrm{d}t}{\alpha - t} = \int_1^{\alpha} \frac{(\alpha - t)^k \mathrm{d}t}{t} = \alpha^k (\ln \alpha - H_k) - \sum_{j \geqslant 1} \binom{k}{j} (-1)^j \frac{\alpha^{k-j}}{j}, \qquad \text{(B.3)}$$

and it remains to evaluate $\int_0^{\alpha-1} \rho_2(t) t^k \mathrm{d}t$ to $\mathrm{O}(\alpha^{k-r} \log \alpha)$. Since $\rho_2 = S_1 + \rho_1$, we have

$$\begin{aligned} \int_0^{\alpha} \rho_2(t) t^k \mathrm{d}t &= \int_0^{\alpha} t^k \mathrm{d}t \left(\int_0^t \frac{\rho(u-1)}{t+1-u} \mathrm{d}u + \rho(t) \right) \\ &= \left(\int_0^{\alpha} \rho(u-1) \mathrm{d}u \int_u^{\alpha} \frac{t^k}{t+1-u} \mathrm{d}t \right) + a_k + \mathrm{O}(\alpha^{-r-1}) \\ &= \int_1^{\alpha} \rho(u-1)(u-1)^k \ln(\alpha + 1 - u) \mathrm{d}u \\ &\quad + \sum_{j \geqslant 1} \binom{k}{j} \frac{1}{j} \int_1^{\alpha} \rho(u)(u-1)^{k-j}((\alpha + 1 - u)^j - 1) + a_k + \mathrm{O}(\alpha^{-r-1}) \\ &= \sum_{1 \leqslant j \leqslant k} \left(\alpha^j - \binom{k}{j} \right) a_{k-j}/j + (\ln \alpha - H_k + 1) a_k \\ &\quad - \sum_{1 \leqslant j \leqslant r} \alpha^{-j} a_{k+j}/j + \mathrm{O}(\alpha^{-r-1}), \qquad \text{(B.4)} \end{aligned}$$

where $a_k = \int_0^{\infty} \rho(t) t^k \mathrm{d}t = A c_k$. Putting all this together and summing leads to the formula

$$S_2(\alpha) = (2 \ln \alpha + 1) \rho_2(\alpha) - \frac{2b_0}{\alpha} - \frac{2b_1}{\alpha^2} - \cdots - \frac{2b_{r-1}}{\alpha^r} + \mathrm{O}(\alpha^{-r-1}), \qquad \text{(B.5)}$$

where

$$b_k = H_k a_k + \sum_{1\leqslant j\leqslant k} \binom{k}{j} a_{k-j}/j. \tag{B.6}$$

In particular,

$$\langle b_0, b_1, b_2, \ldots\rangle = A\langle 0, 2, 19/4, 415/36, 551/18, 13391/150, 1023289/3600, \ldots\rangle.$$

Since $\rho_3 = \frac{1}{2}(\rho_1(\alpha) + \rho_2(\alpha) + S_2(\alpha))$, we have the desired asymptotic series,

$$\rho_3(\alpha) = (\ln \alpha + 1)\rho_2(\alpha) - \frac{b_0}{\alpha} - \cdots - \frac{b_{r-1}}{\alpha^r} + O(\alpha^{-r-1}). \tag{B.7}$$

Incidentally, it can be shown as in Section 6 that

$$\int_1^\infty (\rho_3(t) - e^\gamma(1 + \ln t)/t)\mathrm{d}t = e^\gamma\left(1 + \frac{\pi^2}{12} + \frac{3\gamma^2}{2}\right) - 1. \tag{B.8}$$

References

[1] N.G. de Bruijn, On the number of positive integers $\leqslant x$ and free of prime factors $> y$, *Proc. Kon. Ned. Akad. Wet.* **A54** (*Indag. Math.* **13**) (1951) 50–60.

[2] N.G. de Bruijn, On a function occurring in the theory of primes, *J. Indian Math. Soc.* A **15** (1951) 25–32.

[3] Charles de la Vallée Poussin, Sur la fonction $\zeta(s)$ de Riemann et le nombre des nombres premiers inférieurs à une limite donnée, *Mém. Couronnés Acad. Roy. Belgique* **59** (1899) 1–74.

[4] Karl Dickman, On the frequency of numbers containing prime factors of a certain relative magnitude, *Ark. Mat., Astronomi och Fysik* **22A**, 10 (1930) 1–14.

[5] P. Erdös and M. Kac, The Gaussian law of errors in the theory of additive number theoretic functions, *Amer. J. Math.* **26** (1940) 738–742.

[6] S.W. Golomb, L.R. Welch, and R.M. Goldstein, Cycles from nonlinear shift registers, Prog. Report No. 20–389, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA (1959).

[7] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, 4th ed. (Clarendon Press, Oxford, 1960).

[8] Donald E. Knuth, *Sorting and Searching, The Art of Computer Programming*, Vol. 3 (Addison-Wesley, Reading, Mass., 1973).

[9] William C. Mitchell, An evaluation of Golomb's constant, *Math. Computation* **22** (1968) 411–415.

[10] K.K. Norton, Numbers with small prime factors, and the least k^{th} power non-residue, *Memoirs Amer. Math. Soc.* **106** (1971) 9–27.

[11] J.M. Pollard, A Monte Carlo method for factorization, *BIT* **15** (1975) 331–334.

[12] V. Ramaswami, The number of positive integers $< x$ and free of prime divisors $> x^c$, and a problem of S.S. Pillai, *Duke Math. J.* **16** (1949) 99–109.

[13] L. Shepp and S.P. Lloyd, Ordered cycle lengths in a random permutation, *Trans. Amer. Math. Soc.* **121** (1966) 340–357.

[14] J. van de Lune and E. Wattel, On the numerical solution of a differential-difference equation arising in analytic number theory, *Math. Computation* **23** (1969) 417–421.

[15] J.B. van Rongen, On the largest prime divisor of an integer, *Proc. Kon. Ned. Akad. Wet.* **A78** (*Indag. Math.* **37**) (1975) 70–76.

[16] M.L. Wunderlich and J.L. Selfridge, A design for a number theory package with an optimized trial division routine, *Comm. ACM* **17** (1974) 272–276.